Intelligent Data Analysis date, 2016

Homework number

Professor: Arminda Moreno Díaz

Students: Jose Luis Contreras Santos, Antonio Javier González Ferrer and Alejandro González Pérez

Contents

1 Analysis of the dataset

2 Regression

2.1 Introduction

The aim of this section is to use linear regression to model and predict a wine's quality based on its physicochemical attributes. To do this, we will consider the variable 'quality' as a continuous variable ranging from 0 to 10. It is worth noting, however, that this variable is originally a categorical one which can only take one of the 11 integer values comprised between 0 and 10. Therefore, as predictions will be continuous, many of them will most likely be slightly off due to this discrepancy. This should not pose a problem, as the usual error metrics, such as RMSE, can be obtained anyways.

2.2 Training and test dataset

The data science pipeline often[1] needs to split the original dataset into two smaller pieces: the train and test datasets. If we only evaluate our models in the same dataset, the results will be overestimated (aka overfitting). To provide honest assessments of the performance of the predictive models, we will need to validate the models using a test dataset, a partition that has not been used to build the models in order to avoid bias.

In this case, the test dataset consists of the 20% of the original data (1300 observations) and the training dataset is composed of 5196 data points.

2.3 Model fitting

After splitting the data in training and test sets, the training set was used to fit a linear regression model. At first, all variables were used as explanatory variables for the model:

$$\begin{aligned}(1) \qquad \hat{y} = \beta_0 + \beta_1 * fixed\_acidity + \beta_2 * volatile\_acidity + \beta_3 * citic\_acid \\ +\beta_4 * residual\_sugar + \ \beta_5 * chlorides \\ +\beta_6 * free\_sulfur\_dioxide \\ +\ \beta_7 * total\_sulfur\_dioxide \\ +\ \beta_8 * density + \beta_9 * pH \\ +\ \beta_{10} * sulphates + \beta_{11} * alcohol\end{aligned}$$

However, the marginal coefficient test for each of the variables indicates that the p-values for citric acid and chlorides are not low enough to reject the null hypothesis. Thus, we can consider that the contribution of these to the variance of quality is not significantly greater than 0, and so they have been removed from the final model. The resulting linear regression model is described below:

$$\begin{aligned}(2) \qquad \hat{y} = 59.58 + 0.0616 * fixed\_acidity - 1.275 * volatile\_acidity + 0.04318 * residual\_sugar \\ +\ 0.00618 * free\_sulfur\_dioxide - \ 0.00257 * total\_sulfur\_dioxide \\ -\ 0.5899 * density + \ 0.5068 * pH \\ +0.7161 * sulphates + 0.2637 * alcohol\end{aligned}$$

[1] Some statistical learning models are robust enough to do not need this division. They can infer the behavior of the whole population from a sample if some statistical hypothesis are fulfilled.

According to the obtained model, the two variables with the highest influence on quality are 'volatile acidity' and 'density', specially the latter. This can be misleading however, as it represents the variation in quality per unit of the input, and the scales of the input variables differ in their order of magnitude. Checking the distribution of *density*, for example, the difference between the maximum and minimum values is lower than 0.03, whereas $total\_sulfur\_dioxide$ has a range of over 400. In any case, some conclusions can be extracted from this output. For example, denser wines tend to have a smaller perceived quality due to the negative coefficient, and, on the contrary, those with higher alcohol contents obtain highger quality perception.

The adjusted R-squared coefficient of the resulting model is rather low, with a value of only 0.29. The regression overall p-value, however, is low enough for us to be confident that there exists a relationship between at least part of the input and the output variable, *quality*, performing better than just the simple constant model.

2.4 Assessing the model: assumptions

Let us further examine the quality of the linear regression model by checking if the assumptions are met. In particular, we are interested in testing the linearity of the model and if the residuals are independent, normal and have constant variance (LINE conditions).

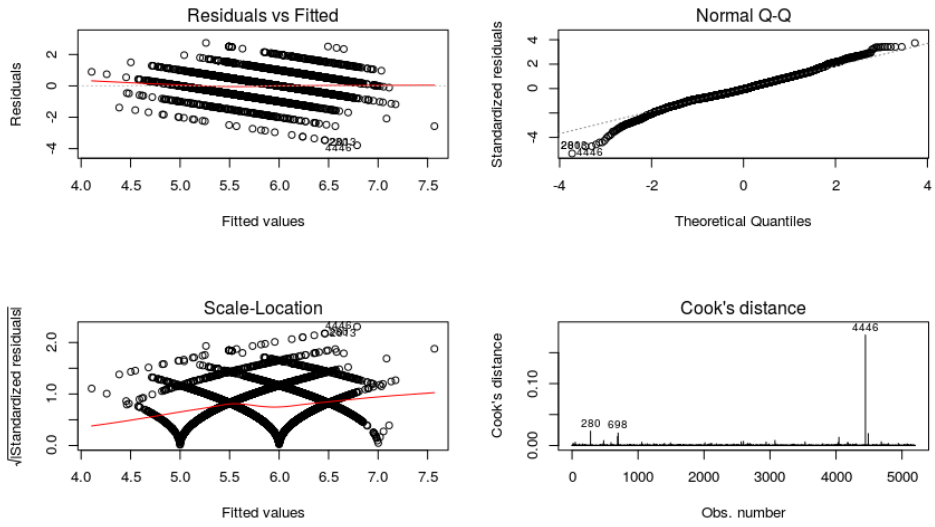

Figure 1: Plot diagnostics for the linear regression model

The conducted tests have different results. First of all, the Rainbow test tests the linear relationship between the response and the linear predictor. The p-value is high enough to not reject the null hypothesis (0.4325) but in other datasets this value is close to 1. On the other hand, the Durbin-Watson test returns a high p-value (0.7377), sign that there is not enough evidence to reject the null hypothesis of autocorrelation of the residuals. Hence, the residuals can be considered as independent. The results for the Jarque-Bera and Breusch-Pagan tests, however, are not so positive. According to their results, residuals are neither normal nor homoscedastic.

Although the residual plots of Figure 1 are unusual due to the clustering around the integer values, they confirm the previous results. The Normal Q-Q plot is in line with the results of the Jarque-Bera test, showing departures from normality, specially in the first quartiles. From the Cook's distance plot, a significant outlier can be seen, corresponding to observation 4446. Even if we remove this outlier, the statistical assumptions remain the same.

3 Logistic Regression

3.1 Introduction

Oh!, a wine factory is going to receive a new pack of different wines and they do not have their type labelled (red or white). Ok, don't worry, we can go through each of the wines, look at it color, and label it. But... we would like to do this process automatically. In the following lines, we will face a classification problem to predict if the wine is red or white, depending on its physicochemical attributes.

A classification problem relates input variables x to the output variable y, but now y can take only discrete values, instead of continuous variables as in regression. When y can only take two discrete, it is called binary classification. We will denote these values as $y \in \{0, 1\}$ in the rest of the report, where $0 \equiv$ white class and $1 \equiv$ red type.

3.2 Models

The equivalent linear regression model in classification is the logistic regression model. This model needs to specify a function such that $p(y = 0|\tilde{\boldsymbol{X}})$ and $p(y = 1|\tilde{\boldsymbol{X}})$ are both greater than 0 and sum 1. The logistic function has such properties, defining the following model:

$$p(y|\tilde{\boldsymbol{X}}, \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}\tilde{\boldsymbol{X}}}}{1 + e^{\boldsymbol{\beta}\tilde{\boldsymbol{X}}}}$$

If $\beta_i > 0$ then increasing one unit in x_i will increase the probability of a success. If $\beta_i < 0$, then the probability of success decrease when increasing x_i. When $\beta_i = 0$, $e^0 = 1$, so the odds do not change with x_i.

3.2.1 Full model

We start by defining a logistic regression model with all the 11 attributes as the predictors. We do not use the 'quality', used in regression, and neither the 'type', used as the target variable y:

$$\begin{aligned} (1) \qquad logit(\hat{y}) = \beta_0 + \beta_1 * fixed\_acidity + \beta_2 * volatile\_acidity + \beta_3 * citic\_acid \\ +\beta_4 * residual\_sugar + \ \beta_5 * chlorides \\ +\beta_6 * free\_sulfur\_dioxide \\ +\ \beta_7 * total\_sulfur\_dioxide \\ +\ \beta_8 * density + \beta_9 * pH \\ +\ \beta_{10} * sulphates + \beta_{11} * alcohol \end{aligned}$$

At first sight, each of the coefficient has a marginal test which attempts the null hypothesis H_0: $\beta_i = 0$, after adjusting the coefficients within the model. That means, it is checked the net effect of each variable and whether should be in the model or not. All the p-values are small enough to reject H_0 (considering $\alpha = 0.05$) except for 'citric_acid' (0.12) and 'sulphates' (0.249). Let us discard these two variables in the further analysis:

$$\begin{aligned} (2) \qquad logit(\hat{y}) = \beta_0 + \beta_1 * fixed\_acidity + \beta_2 * volatile\_acidity + \beta_3 * residual\_sugar \\ +\ \beta_4 * chlorides + \beta_5 * free\_sulfur\_dioxide \\ +\ \beta_6 * total\_sulfur\_dioxide + \ \beta_7 * density \\ +\ \beta_8 * pH + \beta_9 * alcohol \end{aligned}$$

All the marginal tests are now significantly small and the overall fit of the model is high enough ($p = 1$) testing against a Chi-Squared Distribution (pchisq), so we do not have evidence to reject the model in favor of the simple constant model.

We have defined the following metric to validate the correctness of the model based on the confusion matrix of the model. It basically counts the number of correctly classified observations, and it is divided by the total number of examples:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Hence, with the full model we obtain an accuracy of 99% in both train and test dataset. This can be explained by the fact that the 'type' of a wine is clearly defined by a combination of its chemical properties, as expected.

3.2.2 Simpler models

Though we obtain a satisfactory accuracy using almost all the variables of the dataset, we would like to find out a simpler model, where just a few attributes were used. This would lead to a more understandable model, easy to interpret and efficient. For instance, we could agree that an optimal model is the one which provides, at least, a 95% of correct classification.

3.2.2.1 Constant model Let us start with the simplest model: the constant model. Since we know that the classes are a bit unbalanced, let us start with the model which sets all the labels to 1.

$$(3) \qquad logit(\hat{y}) = 1$$

A bit more than 75% of accuracy just by guessing that all the wines will be red. However, we are not using the chemical information. Let us now include one of the variables to the logistic model. Which one? The one which decreases the most the AIC. The AIC is a measure of the quality of different models, relative to each of the other models. Ideal for model selection.

The function 'step' does this task for us: it chooses a model by AIC in a stepwise algorithm. We would use it in the forward direction: tt starts by the simplest constant model, and it tries to achieve the best model up to the full model, previously defined. Since we will go step by step, we will set the number of 'steps' manually, to see what happens in each level.

Variable	AIC
total_sulfur_dioxide	2306.6
volatile_acidity	3522.8
chlorides	3829.2
free_sulfur_dioxide	4178.3
fixed_acidity	4594.0
residual_sugar	4925.6
density	4929.1
pH	5278.9
alcohol	5836.3

Table 1: Comparing logistic regression models using just one predictor

Looking at the output in Table 1, we observe that the best attribute to build a logistic regression with just a single variable is the 'total_sulfur_dioxide'. The accuracy of the logistic regression variable for both train and test datasets is 92%! So finally, 'type' is almost a matter of sulfur in the liquid. This is, nevertheless, not a surprise, since the most correlated variable with respect to the 'type' is also this one (-0.7003).

3.2.2.2 Two variables model (95% accuracy) Let us include one more variable to the following model and see what happens:

(4) $$logit(\hat{y}) = \beta_0 + \beta_1 * total\_sulfur\_dioxide$$

Variable	AIC
density	1300.4
volatile_acidity	1405.4
chlorides	1551.1
fixed_acidity	1942.0
alcohol	2072.2
pH	2142.6
residual_sugar	2279.6
free_sulfur_dioxide	2282.7

Table 2: Comparing logistic regression models using two predictors, given 'total_sulfur_dioxide'.

The AIC decays the most with the inclusion of 'density', reaching an accuracy of 95%. Using just two variables of the dataset, we are just wrong in 58 classifications (21 should have been red, and 37 white) up to a total of 1300 observations.

$$(5) \quad logit(\hat{y}) = -782.71663 - 0.07262 * total\_sulfur\_dioxide + 792.08044 * density$$

The odds can be interpreted such that the 'density' contributes to the probability of getting a red wine (success) while the 'total_sulphur_dioxide' increases the probability of being classified as a whine type, controlling for the other variable.

We stop here, since the addition of more variables does not significantly increase the accuracy of the model. The following plot summarizes the accuracy of the logistic model with respect the number of variables included in the model, following the stepwise algorithm.

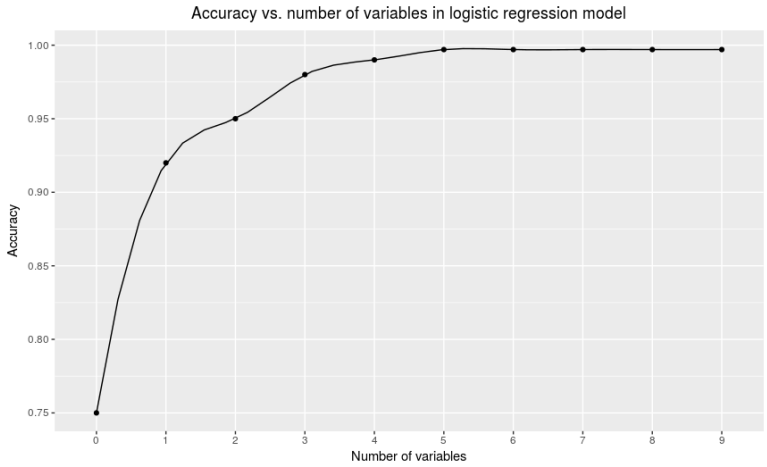

Figure 2: Tradeoff between the number of variables used as predictors in the logistic regression and the accuracy.

3.2.2.3 Five variables model (99% accuracy) We can detect in Figure 2 that using just 5 variables in the logistic regression mode we achieve the same accuracy (99%) than using all the 9 variables from the full model (2), getting this final model:

$$\begin{aligned} (6) \qquad logit(\hat{y}) = -1.945e^{+3} - 4.375e^{-2} * total\_sulfur\_dioxide + 1.937e^{3} * density \\ -7.818e^{-1} * residual\_sugar \\ + 2.065 * alcohol \\ + 7.016 * volatile\_acidity \end{aligned}$$

Both 'total_sulphur_dioxide' and 'residual_sugar' decreases the odds of red wine. The probability of being a red wine is less than the probability of being a white whine if you have high values of these two variables. On the other hand, 'density', 'alcohol'and 'volatile_acidity' increases the odds of being red type. For instance, when 'density' is increased by one unit and all other variables are held constant the odds of $y = 1$ are multiplied by e^3.

3.3 Discussion

After our analysis, we have seen that two different approaches are possible in order to solve our classification problem, introduced at the beginning of this document:

- **Best classification**: In this case, we would like to get the best accuracy as possible. We do not worry in terms of interpretation, but we are looking for the model with less number of parameters that reaches the best accuracy. The model (6) has a 99% of success using 5 of the original variables.

- **Simplest explanation**: We are interesting in understanding the data and interpret the parameters of the classification predictors as straightforward as possible, imposing a minimum of accuracy. In our case, the model (5) reaches a 95% accuracy using just 2 variables, which is very useful to easily describe the model. For instance, we can plot these two variables in the plane:

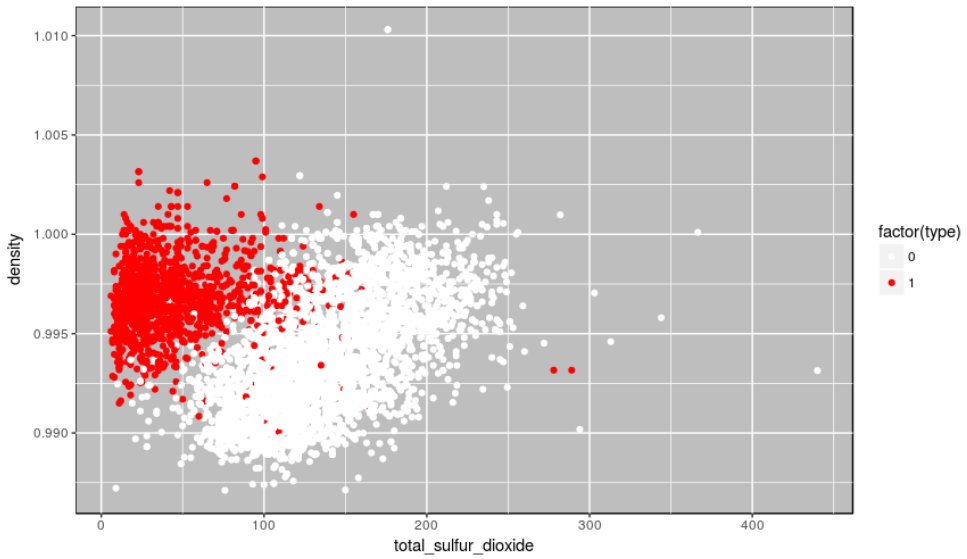

Figure 3: Plot total SO_2 vs. density and colored by type.

This plot shows the variables 'density' and 'total_sulphur_dioxide' but colored by type. As we can see, it is clear that we have an almost perfect clusters between red and white wines. That is why we are having such a great 95% accuracy in our simpler model.

TO (POSSIBLE) DO:

- Analisis de residuos (o comentar porque no los analizamos).
- Hacer plot(model) y comentar algo, especialmente del último plot.
- Another point.